Homework 10

Charles Swarts
Swarts2

April 2016

1 10.7

1.1 question

Lean and Fatty Zucker Rats Zucker rats are specially bred to have curious weight properties, related to their genetics (look them up on the Web). You measure 30 lean Zucker rats, obtaining an average weight of 500 grams with a standard deviation of 50 grams. You measure 35 fatty Zucker rats, obtaining an average weight of 1000 grams with a standard deviation of 100 grams. In steps, you will assess the evidence against the claim that a fatty Zucker rat has exactly twice the weight of a lean Zucker rat. You know that the product of a normal random variable and a constant is a normal random variable. You should assume (and accept, because I won't prove it) that the sum of two normal random variables is a normal random variable.
(a) Write L(k) for the random variable obtained by drawing a uniform sample of k lean rats and averaging their weights. You can assume that k is large enough that this is normal.

- What is E L(k) ? (write an expression, no need to prove anything)
- What is std L(k) ? (write an expression, no need to prove anything)

(b) Now write F(s) for the random variable obtained by drawing a uniform sample of s fatty rats and averaging their weights. You can assume that s is large enough that this is normal.

- What is E F(s) ? (write an expression, no need to prove anything)
- What is std F (s) ? (write an expression, no need to prove anything)

(c) Write popmean(L) for the population mean weight of lean rats, and popmean (F) for the population mean weight of fatty rats. Assume that 2popmean (L) = popmean (F).

- In this case, what is E F (s) - 2L(k) ?
- In this case, what is std F (s) - 2L(k) ?
- Your expression for std F (s) - 2L(k) will have contained terms in the population standard deviation of F and L. What is the standard error of F(s) -2L(k)?

(d) Now assess the evidence against the hypothesis that a fatty Zucker rat weighs exactly twice as much as a lean Zucker rat.

1.2 answer

(a) For this problem, we have at least 30 elements in each sample, therefore we can use the population calculations instead of the sample calculations.

$$\mathbb{E}(L^k) = popmean(L) \approx L^k = 500\text{grams}$$

$$std(L^k) = \frac{popstd(L)}{\sqrt{k}} \approx \frac{std(\mathbf{l})}{\sqrt{k}} = \frac{50}{\sqrt{30}} = 9.12870929$$

(b)

$$\mathbb{E}(F^s) = popmean(F) \approx F^s = 1000\text{grams}$$

$$std(L^k) = \frac{popstd(L)}{\sqrt{k}} \approx \frac{std(\mathbf{f})}{k} = \frac{100}{\sqrt{35}} = 16.9030850$$

(c) For this question, the $std(F^s - 2L^k)$ is the standard error in the experiment we are considering.

$$\mathbb{E}(F^s - 2 * L^k) = \mathbb{E}(F^s) - 2\mathbb{E}(L^k) = 1000 - 2 * 500 = 0$$

$$std(F^s - 2L^k) = \sqrt{(std(F^s))^2 + (std(2 * L^k))^2} = \sqrt{(std(F^s))^2 + 4 * (std(L^k))^2}$$

$$= \sqrt{\left(\frac{popstd(F)}{\sqrt{s}}\right)^2 + 4 * \left(\frac{popstd(L)}{\sqrt{k}}\right)^2}$$

$$\approx \sqrt{\left(\frac{std(\mathbf{f})}{\sqrt{s}}\right)^2 + 4 * \left(\frac{std(\mathbf{l})}{\sqrt{k}}\right)^2} = \sqrt{\left(\frac{100}{\sqrt{35}}\right) + 4 * \left(\frac{50}{\sqrt{30}}\right)^2} = 24.88066757$$

(d) The formula for evaluating the evidence is

$$\frac{\text{result} - \text{expected}}{SE}$$

$$= \frac{0 - 0}{24.88066757} = 0$$

You see, it's a mute point because we hypothesised that the mean Fat Zucker rat would be twice as heavy as the mean Lean Zucker rat, and the data give met exactly that description, so the result-expected term goes to zero. That suggests that indeed the hypothesis is true. We would expect to get a result like this or within one standard deviation of this 68% of the time. Also, 50% of the time, the value will be greater and 50% of the time, the value will be less.

2 10.10

2.1 question

10.10. Assume the average weight of an adult male short-hair house cat is 5 kg, and the standard deviation is 0.7 kg (these numbers are reasonable, but there's quite a lively fight between cat fanciers about the true numbers).
(a) What fraction of samples consisting of 30 adult male short-hair house cats (selected uniformly at random, and with replacement) will have average weight less than 4kg?
(b) What fraction of samples consisting of 300 adult male short-hair house cats (selected uniformly at random, and with replacement) will have average weight less than 4kg?
(c) Why are these numbers different?

2.2 answer

(a).

$$\frac{\text{result} - \text{expected}}{SE} = \text{Z-score}$$

$$SE = \frac{SD}{\sqrt{k}} = \frac{.7}{\sqrt{30}} = 0.127802$$

$$\frac{4-5}{0.127802} = -7.824603$$

now I plug that into r using the following command

```
pnorm(7.824603,lower.tail = FALSE)
```

And I get that only $(2.546307e-17)\%$ of the time will I see a sample average of 30 cats be less than 4kg.

(b).

$$\frac{\text{result} - \text{expected}}{SE} = \text{Z-score}$$

$$SE = \frac{SD}{\sqrt{k}} = \frac{.7}{\sqrt{300}} = 0.04041452$$

$$\frac{4-5}{0.0404145} = -24.743$$

now I plug that into r using the following command

```
pnorm(24.743,lower.tail = FALSE)
```

And I get that only $(1.843849e-137)\%$ of the time will I see a sample average of 300 cats be less than 4kg.

Since there aren't even enough cats that have ever lived in general to fulfill that requirement, I'm going to say the probability is practically 0%

(c) These numbers are different because in order to get a group of 300 cats with an average of 4kg, you need a lot more abnormal cats in your group than if you want to distort a group of 30 cats.

3 11.3

3.1 question

You have a dataset x of N vectors, x_i, each of which is d-dimensional. Assume that Covmat (x) has one non-zero eigenvalue. Assume that x1 and x2 do not have the same value. (a) Show that you can choose a set of ti so that you can represent every data item xi exactly xi = x1 + ti(x2 x1). (b) Now consider the dataset of these t values. What is the relationship between (a) std (t) and (b) the non-zero eigenvalue of Covmat (x)? Why?

3.2 answer

Freshmen didn't read the prerequisites. :(

Can't do any linear algebra.

4 11.5

4.1 question

Take the wine dataset from the UC Irvine machine learning data repository at http://archive.ics.uci.edu/ml/datasets/seeds.
(a) Plot the eigenvalues of the covariance matrix in sorted order. How many principal components should be used to represent this dataset? Why?
(b) Construct a stem plot of each of the first 3 principal components (i.e. the eigenvectors of the covariance matrix with largest eigenvalues). What do you see?
(c) Compute the first two principal components of this dataset, and project it onto those components. Now produce a scatter plot of this two dimen- sional dataset, where data items of class 1 are plotted as a '1', class 2 as a '2', and so on.

4.2 answer

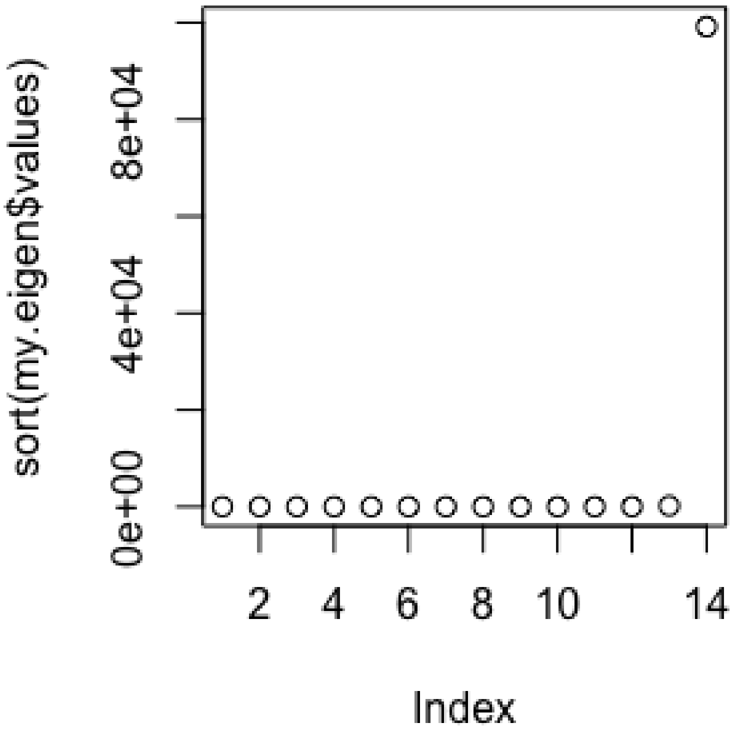
(a) I believe that only the one that looks super important is necessary. The rest have really tiny values comparatively. My final answer is one. (b) I see that even amongst the top three contendors, the first eigenvalue is still the dominant force. This further implies that it is the only necessary component to describe the dataset. (c) I apologize for not doing c.

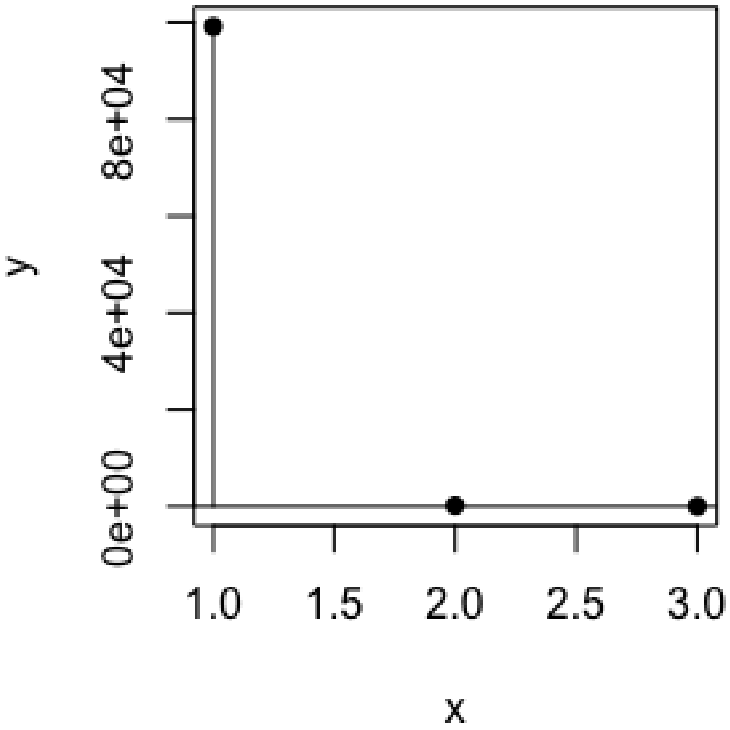
Here is the code which I used to produce the graphs below.

```
    data <- read.csv("/Users/CharlesBSwarts/Desktop/wine.csv")
names <- c("class","Alcohol","MalicAcid","Ash","AlcalinityOfAsh","Magnesium","TotalPhenols","Flavanoid
head(data)
colnames(data) <- names
head(data)
my.cov <- cov(data)
my.cov
my.eigen <- eigen(my.cov)
my.eigen
stem(my.eigen$values)

plot(sort(my.eigen$values))

#The function
stem <- function(x,y,pch=16,linecol=1,clinecol=1,...){
  if (missing(y)){
    y = x
    x = 1:length(x) }
  plot(x,y,pch=pch,...)
  for (i in 1:length(x)){
    lines(c(x[i],x[i]), c(0,y[i]),col=linecol)
  }
  lines(c(x[1]-2,x[length(x)]+2), c(0,0),col=clinecol)
}
stem(my.eigen$values)
```


sort(my.eigen$values)
0e+00
4e+04
8e+04
2
4
6
8
10
14
Index



5 11.6

5.1 question

Take the wheat kernel dataset from the UC Irvine machine learning data repository at http://archive.ics.uci.edu/ml/datasets
Compute the first two principal components of this dataset, and project it onto those components.
(a) Produce a scatterplot of this projection. Do you see any interesting phe- nomena?
(b) Plot the eigenvalues of the covariance matrix in sorted order. How many principal components should be used to represent this dataset? why?

5.2 answer

(a) So sorry

(b) I believe that based on the graph, the top three definitely need to be used because they start to flatten out in importance at the fourth eigenvalue. I would personally include the fourth because it still is a factor of ten more important than the fifth.

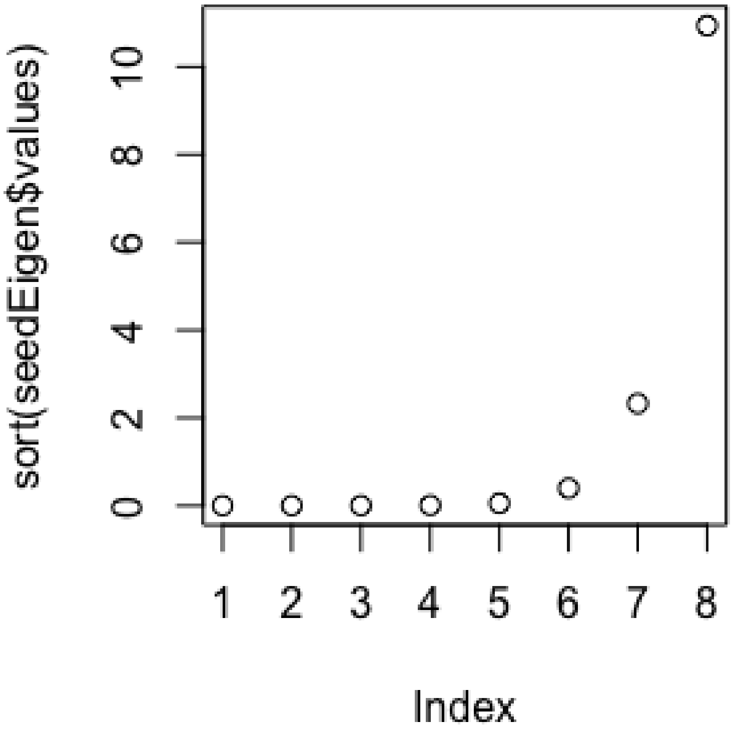
Here is the code that produced the graph below.

```
    seedNames <- c("areaA","perimeterP","compactness","lengthOfKernel","widthOfKernel","asymmetryCoeff
seedData <- read.csv("/Users/CharlesBSwarts/Desktop/seed.csv")

colnames(seedData) <- seedNames

seedCov <- cov(seedData)
seedEigen <- eigen(seedCov)

plot(sort(seedEigen$values))
```


sort(seedEigen$values)
Index